BeNeLux Bioinformatics Conference – Antwerp, December 7-8 2015
Abstract ID: **000** Category: Abstract template

REPRESENTATIONAL POWER OF GENE FEATURES FOR FUNCTION PREDICTION

Konstantinos Pliakos[1*], *Isaac Triguero*[2,3], *Dragi Kocev*[4], *Celine Vens*[1].

Department of Public Health and Primary Care, KU Leuven Kulak[1]*; Department of Respiratory Medicine, Ghent University*[2]*; Data Mining and Modelling for Biomedicine group, VIB Inflammation Research Center*[3]*; Department of Knowledge Technologies, Jožef Stefan Institute*[4]*;*
[*]*konstantinos.pliakos@kuleuven-kulak.be*

We present a short study on gene function prediction datasets, revealing an existing issue of non-unique feature representation, as well as the effect of this issue on hierarchical multi-label classification algorithms.

INTRODUCTION

This study focuses on hierarchical multi-label classification (HMC). HMC is a variant of classification where one sample can be assigned to several classes simultaneously. It differs though from multi-label classification as these classes are organized in a hierarchy. That means that a sample belonging to a class automatically belongs to all its super-classes. Typical HMC tasks include gene function prediction or text classification. Here, we focus on the former.

A typical characteristic of genes is that they can be described in several ways: using information about their sequence, homology to well-characterized genes, expression profiles, secondary structure of their derived proteins, etc. The HMC community has multiple research datasets at its disposal on gene functions (e.g., (Vens *et al.*, 2008) or (Schietgat *et al.*, 2010)), each representing genes by one type of features. Indisputably, researchers should get advantage of this amount of data but the question arises how "good" these datasets are. How discriminant are the features describing a gene? Here, a short study is trying to display existing data-related problems and give answers to the aforementioned questions.

DATA STUDY & RESULTS

After careful experimentation on various publicly available datasets it was noted that some of them suffer from large amount of duplicate feature vectors. The irrational behind this occurrence is that there are genes, which despite having different functions, have exactly the same feature representation. The table below lists the aforementioned problem in the 20 gene function prediction datasets described in (Vens *et al.*, 2008) and (Schietgat *et al.*, 2010).

Organism	Dataset	Nb of genes	Nb of unique gene representations
S. cerevisiae	church	3755	**2352**
	pheno	1591	**514**
	hom	3854	3646
	seq	3919	3913
	struc	3838	3785
A. thaliana	scop	9843	9415
	struc	11763	11689

TABLE 1*. Datasets, the number of genes and their unique representations.*

As it is displayed, the church (micro-array expression) and the pheno (phenotype features) datasets suffer the most. More specifically, in pheno dataset the 67.7% of the gene representations are duplicates. The most frequent feature vector appears 315 times, 197 times in the training set and 118 times in the test set. Due to this, 20% of the 582 test examples will give the same feature vector as input for prediction. In a decision tree model, for example, these genes will end up in the same leaf, receive the same prediction (the average class vector of 197 training examples), but receive a different error term as they are a priori associated with a different class label-set. In the training phase, there may still be a lot of variation in the class vectors of the 197 genes, but no split exists to separate them. In the Church dataset, the 3755 genes correspond to only 2352 unique feature descriptors. In Hom or Struc datasets the number of the duplicates is lower but still impressive, considering the enormous size of the feature vectors in these datasets.

For evaluation purposes, ML-KNN (Zhang M. L *et al.*, 2007) was employed to demonstrate the effect of the studied problem on the average precision for the FunCat annotated datasets. Here, "unique" refers to the datasets occurring after removing all the duplicates. Thus, any feature vector can only once be included in a gene's neighbour set. We report the average of 10 "unique" versions, each one using a different gene's class label as ground truth for the feature vector.

Dataset		K= 1		K = 5		K = 17	
		Train	Test (5cv)	Train	Test (5cv)	Train	Test (5cv)
pheno	initial	51.59	23.62	39.55	24.14	32.76	23.59
	unique	**100**	**24.21**	**55.62**	**24.90**	**39.70**	**25.01**
hom	initial	98.30	39.32	63.64	39.45	48.96	37.28
	unique	**100**	39.14	**64.64**	**39.67**	**49.28**	**37.53**

TABLE 2*. Average Precision rates (%) using ML-KNN.*

The table shows that the less discriminant feature representation can affect the ML-KNN and decrease the precision of multi-label classification. Indisputably, it could be concluded that the same problem will be more obvious or even completely disastrous for two-class or multi-class classification problems.

CONCLUSION

The major point of this study was to inform the research community of the relatively low representational power of the features present in some widely used gene function prediction datasets, making them even more difficult and challenging datasets from machine learning perspective. We observed the same issue in datasets of other HMC application domains like text categorization.

REFERENCES

Zhang M. L. & Zhou Z. H. ML-KNN: A lazy learning approach to multi-label learning, *Pattern recognition* **40**, 2038-2048, (2007).

Vens C. *et al.* Decision trees for hierarchical multi-label classification, *Machine Learning* **73,** 185-214, (2008).

Schietgat L. et al. Predicting gene function using hierarchical multi-label decision tree ensembles, *BMC Bioinformatics* **11**, (2010).